# Instuderingsfrågor

Frågorna kommer diskuteras i grupper och därefter diskuteras i helklass. Dessa frågor skall hjälpa oss i vår inlärning och du förväntas inte kunna alla svar.

## Frågor kopplat till slides från presentation 1.

1. Hur kan vi definiera maskininlärning?
2. Vad kännetecknar regressionsproblem?
3. Vad kännetecknar klassificieringsproblem?
4. Antag att vi har följande regressionsmodell: , vad kallas Y, X, och ?
5. Kan du ge ett exempel där följande regressionsmodell hade kunnat användas: ?
6. Hur kan vi utvidga/generalisera modellen i fråga 5?
7. När man skall göra modellval och modellutvärdering så kan man använda ett tränings, validerings och test data. Förklara hur de olika delarna används.
8. Om man vill köra K-fold cross validation så kan man börja med att dela in datan i träning och test set, förklara hur K-fold cross validation fungerar.

## Frågor kopplat till slides från presentation 2.

1. Kan du ge 2 exempel på vad klassificeringsmodeller kan användas till?

De kan användas för att beräkna churn, dvs. vilka kunder som har lämnat företaget eller inte, och e-postfiltrering, t.ex. spam eller inte spam.

1. Hur tolkar du följande Confusion Matrix:   
   En bild som visar text, skärmbild, Teckensnitt, nummer

   Automatiskt genererad beskrivning
2. 165 patienter analyserades, varav 5 räknades som falskt negativa eller FN och 10 som falskt positiva eller FP. Det innebär att 5 av 165 bedömdes vara vid liv men dog så att de inte fick den vård de har laglig och moralisk rätt till.
3. Hur definieras Precision? Hur tolkas Precision?  
   Precisionen baseras på noggrannheten i modellen som mäter positiva förutsägelser som är korrekta, så den definieras som ett förhållande mellan sanna positiva förutsägelser (TP) över totala positiva förutsägelser (TP + FP).
4. Hur definieras Recall? Hur tolkas Recall?  
   Recall baseras på modellens förmåga och känslighet och tittar på vad vi förutspådde som korrekt när det gäller verkliga data. Det definieras som ett förhållande mellan True Positive-förutsägelser (TP) över Total Positive-data (TP + FN).
5. Hur kan binära klassificerare användas för ”multiklass klassificierings problem”?   
   Ledning: OvO och OvR algoritmerna.

Du separerar de binära klassificeringarna i olika klasser eller grupperingar antingen 1 på 1 eller 1 till många och itererar sedan genom varje klass.

1. Hur funkar OvO och OvR algoritmerna?   
   OvO trains the classes in a multiclass classification on data from two other classes, which it iterates through based on a quadratic function.
2. OvR takes each object and sets it against all the rest of the objects and then iterates through each of these as individual groups. classifier is responsible for distinguishing between one of the classes and all other classes combined.
3. Vad innebär ”Multioutput-Multiclass Klassificiering”?

Detta innebär att du har flera underkategorier inom varje klass.

## Frågor kopplat till slides från presentation 3.

1. En bild som visar Teckensnitt, diagram, vit, text

   Automatiskt genererad beskrivningEn bild som visar Teckensnitt, skiss, vit, linje

   Automatiskt genererad beskrivningHur kan du skriva ut följande summor:

1 + 2 + 3 + 4 + 5 = 15 och 9 + 4 + 1 = 14

1. En bild som visar Teckensnitt, diagram, vit, linje

   Automatiskt genererad beskrivningRoot Mean Squared Error (RMSE) är ett viktigt mått för att utvärdera regresionsmodeller:

Hur kan du tolka måttet?

This shows the average of the squares of each value from 1 to n observations of the difference between actual and predicted values.

Detta visar medelvärdet av kvadraterna för varje värde från 1 till n observationer av skillnaden mellan faktiska och förutsagda värden.

1. Kan du kort förklara hur fixa men okända parametrarna , skattas med hjälp av MSE?

The fixed but unknown parameters *β*0​,*β*1​,…,*βp*​ are estimated by minimizing the MSE between the actual and predicted values. This can be done using the Ordinary Least Squares method, which finds the parameter values that best fit the model to the data by minimizing the sum of squared differences between observed and predicted values.

De fasta men okända parametrarna β0,β1,...,βp uppskattas genom att minimera MSE mellan de faktiska och förutsagda värdena.   
Detta kan göras med hjälp av den Ordinary Least Squares method, som hittar de parametervärden som bäst passar modellen till data genom att minimera summan av kvadratiska skillnader mellan observerade och förutsagda värden.

1. I frågan ovan, varför används MSE och inte RMSE? Vad är skillnaden?

RMSE omvandlar MSE tillbaka till samma skala som användes av de observerade och förutsagda värdena.

1. Vad är intuitionen bakom ”Gradient Descent”, vad används det till?

Gradient Descent works based on the idea of iterating towards the best function and is used often in linear regression to find the best cost function for each parameter.

1. Vad är en parameter? Vad är en hyperparameter? Vad heter dem två koncepten i Scikit-learn?

In creating a model, a parameter is a constant that is estimated using the training data. A hyperparameter is used to adjust the parameters before training the model. In Scikit-Learn a HP is called a Parameter and a Parameter is called an Attribute.

1. Vad är Bias Variance Trade-off? ¨

This means that you try to balance the inherent bias of simple data models with little noise versus the need for more and more varied data sensitivity in complex models, but which catch more noise.

1. Varför är mer komplexa modeller inte alltid bättre?

Because they are prone to overfitting the model on training and thus perform poorly on test data.

1. Vad är intuitionen bakom att regularisera en modell? Har du tre exempel på regulariserade regressionsmodeller?

Regularize a model aims to by introduce a penalty term to the loss function used to train the model, discouraging it from overfitting.

1. Varför är namnet för ”Logistisk Regression” missvisande?

För att det använder för att lösa klassification problem inte regression.

1. A graph of a function

   Description automatically generatedVarför används ”Logistic Function”? Se slide 40 – 44. ¨

The logistic function is used in logistic regression, where it models the probability that a given weighted input (plus baseline intercept) belongs to a particular category. In a plot, over the 50% threshold log-odds values on x-axis belong to positive class and under this threshold belong to the negative class.   
The higher up or down the curve the more likely category belongs to a particular class.

The steepness of the curve shows how quickly predictions move from one class to the other as the input value changes. A steeper curve means a small change in x can significantly affect the predicted probability.

Den logistiska funktionen används i logistisk regression, där den modellerar sannolikheten för att en viss indata tillhör en viss class.